Combining Environmental Information

Walter W. Piegorsch
Department of Statistics, University of South Carolina, Columbia, SC 29208 USA

An increasingly important concern in environmental studies is the need to combine information from diverse sources that relates to a common endpoint or effect. Such an activity is statistical in nature, and statistical techniques are integral to analyses that combine such information. These techniques are only developmental, however: modern statistical methods for combining environmental information require subject-specific formulations.

To facilitate further development of techniques in combining information, a workshop was held on 27–28 September 1993 at the University of North Carolina, Chapel Hill, to bring together environmental scientists and statisticians working in the area of collection and analysis of environmental data. The workshop was cosponsored by the U.S. Environmental Protection Agency and the National Institute of Statistical Sciences (NISS). The workshop co-organizers were Lawrence Cox of the U.S. EPA and Jerome Sacks of NISS. At the basic level, the workshop explored statistical problems in combining information as posed by applications in the environmental sciences. Additionally, the workshop identified and disseminated methods and research themes leading to solutions for these problems and stimulated further interdisciplinary collaborations in this rich research area. In effect, the workshop's goals were to inform environmental scientists of relevant statistical methodologies for combining information, expose statisticians to these important environmental applications, and highlight substantive quantitative questions that require attention.

The workshop was organized around a selection of environmental projects and studies that demonstrate the need for combining environmental information. The program covered the following environmental data problems:

- Combining environmental data from multiple and diverse sources. Examples included statistical reporting on environmental conditions and trends in aquatic, terrestrial, and atmospheric settings and combining designed environmental study data and observed environmental assessment data.
- Forming environmental indicators and indices, including combined mapping procedures and multiple data source conformance.
- Combining environmental epidemiologic studies for hazard identification and risk assessment. Emphasis was directed toward public health issues in assessing exposures to environmental tobacco smoke, dioxin, and nitrogen dioxide, assessing acute inhalation risk, assessing effectiveness of lead abatement strategies, and using prior quantitative information and "Bayes" methods to model uncertainty in effect estimation.

Each presentation session included one or two scheduled discussants and open-floor discussion. Participants included environmental and statistical scientists with familiarity and expertise in diverse areas. This remainder of this report briefly summarizes the main themes and issues raised in each of the topic areas at the EPA/NISS Workshop.

Combining Data from Multiple Sources

The first workshop session, "Combining Ecological Assessments," was motivated by quantitative issues identified by EPA's recently instituted Environmental Monitoring and Assessment Program (EMAP). EMAP's goals are to estimate the status of the nation's ecological resources, interrelate these estimates with potential impacts on organisms inhabiting the associated ecosystems (including humans), and provide periodic updates and assessments of these resources. (The latter issue is essentially a concern with trends over time and space. As was noted several times during the workshop, temporal and spatial concerns dominate statistical issues in combining environmental information.)

The featured speakers on combining ecological assessments, R. Linthurst and A. Olsen of the EPA, had collaborated previously on EMAP concerns and provided an excellent overview of the problem. For instance, EMAP's basic biological questions—the state of various ecosystems throughout the United States—lead naturally to combination of information and data from multiple ecological sources. The potential exists for many forms of statistical interaction among these sources, and statistical designs and analyses that account for such interactions must be constructed carefully. Combining data across diverse sources requires knowledge of 1) how each site was selected, 2) the exact definition/nature of the attribute(s) measured, 3) the sampling period and whether any temporal effect(s) may be present, and 4) whether the data garnered from the source is of sufficient quality to provide useful inferences.

The EMAP paradigm is essentially a complex sample survey of ecological population characteristics. In the simplest setting, straightforward statistical methods are available for effect estimation *(1)*. When spatial and temporal effects may bias statistical outcomes from such studies, however, more complex statistical sampling designs are required. The workshop presentations noted that EMAP uses a form of design known as probability-based sampling *(2)*, but that data from other collection programs certainly will be combined with probability-sampled EMAP data. Thus, open statistical questions exist on how to perform data combination from mixed sample designs *(3)*. The associated problems in multiframe/multistage sampling provide many open venues for future research.

Discussion on these issues, led by N. Cressie of Iowa State University, included the reminder that ecological science is organism-based, whereas environmental science typically directs attention to conditions surrounding the organism. Any investigation into either an organism's status in an ecosystem or into the environment affecting that organism and ecosystem will involve other organisms and other ecosystems. Complex interactions are at work in both areas, and the results of any statistical investigation will have broad applications. Indeed, an important point made during the discussion was that although EMAP has derived great motivation from ecological problems, many issues in public health require similar attention: complex sample surveys of human populations can take on related forms, and information from EMAP surveys may provide public health researchers with important input into potentially

Thanks are due to Jerome Sacks, Lawrence Cox, and an anonymous reviewer for helpful suggestions during the preparation of this report. This work is part of cooperative agreement no. EPA-CR-819638-01-0 between the U.S. EPA and the National Institute of Statistical Sciences. The contents of this report are the responsibility of the author; views do not necessarily reflect those of EPA or NISS. This report has not been subjected to EPA's peer or administrative review.

detrimental effects on human populations (and vice versa). Participants agreed that EMAP scientists and public health officials may have to combine their future studies and goals to achieve useful, cost-effective conclusions. (This recognition helped set the stage for a later set of presentations on epidemiological data combination; see below.)

Formation of Environmental Indices

A second workshop session focused on how to formulate and measure indices of environmental damage, particularly as motivated by EPA concerns. For example, EMAP is faced with the task of combining regional resource data, historical data, administrative data, etc. Incorporating such disparate information into a concise, integrated resource assessment is a complex task. A call was made by E. Hyatt of the EPA to convert disparate environmental information into common metrics. Specific questions for statistical consideration were 1) how to construct confidence limits on indices from disparate sources, 2) how to prioritize different indices and scale greater-quality indices to give them more 'weight' in the combined analysis, 3) if and how a general, combined index of environmental indicators can be developed, and 4) whether such an index or multiple indices can be simplified for use by the nontechnical population, decision-makers, etc. The ensuing discussion, led by J. Rawlings of North Carolina State University, recognized that a critical trade-off in any such endeavor pits simplicity against full information retrieval. Loss of information is an anathema to statisticians, but an inability to interpret overly complex measures is just as problematic. Nonetheless, a single, simplified environmental indicator was thought to be unattainable due to the highly complex inputs, interactions, and responses seen in any study of environmental or ecological response to pollutants. Opportunities do exist, however, for productive collaborations on index development among environmental scientists and statisticians. Issues that will arise in these interactions include: 1) definition of appropriate reference populations/systems, 2) clear understanding of the structure of the ecosystem under study, 3) adjustments for local temporal changes (wet versus dry seasons, etc.) that still preserve the effect under study, 4) optimization of spatial and temporal data to obtain indices, 5) standardization of indices when possible, and 6) calibration/recalibration to monitor the operating characteristics of the statistical measure.

Indeed, these six points are as much a specific set of prerequisites for index development as they are a general set for consideration in many of the project areas for combining environmental information discussed herein.

Criteria and Dose–Response Assessment Methodologies

Basic statistical research into combined data analyses over multiple studies has included applications in ecotoxicology *(4)*, business/economics *(5)* and other social sciences *(6)*, and biomedical settings *(7–9)*. Indeed, the commonly coined term for the method is "meta-analysis," although statistical methods for combining information extend beyond this single area of research. This was recognized by the chair of the next workshop session, I. Olkin of Stanford University, a major contributor to research in meta-analysis *(10)*. The session focused on combining information in epidemiology and environmental medicine. An important and topical example of such concerns was illustrated by the first speaker, S. Bayard of the EPA, with a discussion of the methods used for combining epidemiologic data in the recent EPA study on health effects of environmental tobacco (passive) smoke (ETS). The EPA pronouncement in December 1992 that respiratory health effects of ETS can include cancer drew great media attention. This publicity veiled, however, an extensive and complex analysis of data from multiple sources. Thirty epidemiologic studies were considered as part of the EPA analysis, including within-country and between-country combinations using the highest-quality data available. Important statistical innovations included relative risk models that adjusted for background exposures and for potential systematic downward bias due to control group exposures to ETS. Pooled population estimates suggested values as high as a 59% increase risk for lung cancer mortality in U.S. nonsmokers due to ETS exposure. A similar analysis of lung cancer risk after occupational exposure to dioxin illustrated additional models and analyses for modeling risk after environmental exposure.

In a short discussion of the ETS study, the discussion leader, D. Gaver of the Naval Postgraduate School, emphasized that in epidemiologic studies with diverse populations (e.g., multiple countries or regions) subjects often exhibit excess person-to-person variability. This sort of extra variability (or "overdispersion") is an important concern: it must be incorporated carefully in the statistical analysis or incorrect inferences can result *(11)*. Also, improper data standardization among studies can lead to excess study-to-study variation. Here again, the analyst must make certain that data standards are made uniform or calibrated across studies to avoid unexpected overdispersion and incorrect inferences.

The session continued with additional presentations on environmental epidemiologic studies, each calling for greater statistical research into data combination and consequent analyses. A summary by D. Kotchmar of EPA on an EPA meta-analysis of respiratory damage after indoor exposure to nitrogen dioxide highlighted the ability of meta-analytic techniques for synthesizing diverse outcomes and assessing study-to-study similarity [see Hasselblad et al. *(7)*]. The results suggested that each increase of 0.015 ppm NO_2 exposure can lead to an increased risk of respiratory illness of as much as 20% over unexposed controls. A different perspective was presented, however, in an EPA study on environmental lead abatement in reducing children's blood-lead levels, presented by A. Marcus of the EPA. The project involved multiple sites, requiring information combination of various sorts. Specific statistical techniques of interest included multivariate analysis to account for repeated measures on study subjects *(12,13)* and structural equation modeling *(14)* to account for different lead pathways into the bloodstream. The preliminary results showed only minimal success in achieving exterior lead abatement and highlighted problems in meta-analyses of this sort. Statistical problems included certain unwelcome sensitivities to model specifications (a lack of robustness) and difficulty in developing proper lead pathway models for analysis.

A third perspective on the need for new statistical formulations was provided in a report on EPA's development of methodology for acute inhalation risk assessment by D. Guth of the EPA. The project concerned combination of data from studies on inhalation damage from various airborne toxins in order to estimate human health risk. The data varied greatly in their endpoints: short- and long-term exposures in laboratory animals, acute exposures to humans in chemical and/or community accidents, chronic exposure studies in urban areas, etc. Hence, major statistical concerns were raised regarding the type, quantity, quality, and relevance of these data. Current research in this area is focused on data in categorical form, although other forms are possible. The research goal is to develop methodology for data combination that correctly includes the range of endpoint severities and of exposure concentrations and duration.

All these presentations provided important criteria and motivation for development of quantitative methods for data combination and analysis in environmental health settings. To address some of these

issues and to raise concerns about others, the session also provided opportunities for presentations on current statistical approaches for these problems. For instance, a presentation on use of benchmark doses for risk assessment and its statistical implications by V. Hasselblad of Duke University provided an excellent springboard for discussion on statistical methods for parameter estimation over many data sources. The specific form of statistical methodology highlighted was Bayes analysis, where variation over multiple sources is incorporated via a mathematical, prior distribution function *(15)* . The analysis can incorporate uncertainty in the response measure for combining response data from individual studies. Specific attention was directed at improved estimation of benchmark doses and no-observed-effect levels from studies of noncancer endpoints. For a given environmental stimulus of interest, the method allows for statistical combination of data across endpoints *(16)*. The discussion that followed, led by W. DuMouchel, considered a number of complementary statistical approaches, including a novel suggestion of weighted linear regression *(17)* to mimic the data combination effect by viewing it as a heterogeneous variance setting or use of standard random effects models *(18,19)* to incorporate differential effects of the data combination over multiple studies. Random effects analyses are quite similar in nature to Bayes analyses for many statistical models *(20–22)* . In practice, the propriety of one method over the other may be determined by which is most easily implemented given the computer resources available to the analyst. Indeed, a modern interactive computing environment is almost essential for the analysis of complex environmental data, and those methods that take best advantage of the full range of modern exploratory statistical graphics *(23–26)* will be those first considered for use.

A second discussion, led by R. Carroll of Texas A&M University and D. Simpson of the University of Illinois, centered on estimation of benchmark dose and other measures of environmental chemical toxicity. An additional model that incorporated random effects variability across studies was presented based on logistic regression *(27)*. The method was shown to allow for estimation of various parameters of interest, including the benchmark dose, the median effective dose (ED_{50}), etc. Some caveats were raised, however, including the perhaps obvious concern that any method of analysis must possess proper motivation from the subject matter: is a true response curve approximately linear or linear-logistic at low doses, or is there a nonlinear threshold effect? Indeed, how one would combine studies with different response curves was cited as an important, unsolved area of statistical research for environmental applications. Also raised were concerns about proper study design, including number and spacing of doses, when preparing studies of this sort. Some research has appeared on design considerations in these areas (*28–30*), but it was argued that much more remains.

Combining Environmental Data for Statistical Reporting

The next workshop session on hazard identification concerned statistical reporting of results. An overview of these concerns by B. Nussbaum of the EPA identified the issue as a natural one for statistical study, especially when results from previous studies are used to determine if or how further study of an environmental hazard is to be performed. Proper statistical reporting can obviate the need for further studies, saving resources for other projects.

The first presentation of this session, by B. Sinha of the University of Maryland, discussed a statistical method for combining data from several locations within one single site, such as a Superfund waste site. The method was formulated as a combination of *F*-statistics from multiple tests of hypotheses at each site, with an adjustment to correct for multiplicity. A special case of this application was presented next by N. Nagaraj of the University of Maryland and R. Shafer of the EPA. At issue was assessing nutrient loading due to pollution in Chesapeake Bay. (The study is in the planning stages.) A benthic index was devised to measure the pollutant effects and will be used by multiple agencies studying the bay. A mapping procedure was constructed to incorporate relevant aspects of the aquatic community that contribute to the index. The mapping must be as robust as possible to spatial variations; to achieve this a hybrid analysis was proposed, using features of robust multiple regression *(31,32)* and a form of data combination known as kriging *(33,34)* . Enhanced ability to report the data for public use is anticipated.

The session also included a presentation on international standards for data reporting by B. Bargmeyer of the EPA. Critical to the use and sharing of environmental data is the need to set unambiguous standards for naming, defining, and documenting data elements, yet the fundamental principles of such data representation are barely in place. It was noted that EPA is working with national and international standards bodies to establish these data standards.

Discussion on all these issues in statistical reporting, led by K. Reckhow of Duke University and D. Carr of George Mason University, emphasized the basic needs: standard forms for plotting data, recognizing spatial effects, and identifying correct statistical features of the spatial variability. Guidance was called for from subject-matter scientists working in appropriate environmental areas, especially where data quality varies. Participants were reminded that spatial features of water sources such as the Chesapeake Bay can be deceptive and may be improperly modeled and analyzed. Bayes-type models and meta-analyses (see above) can be useful here, but robustness and/or resilience of any method to unrecognized spatial (or other) variations is a critical characteristic. Further study is necessary, and this promises to be an important and active area of future interdisciplinary, environmetric research.

Summary

Workshop proceedings and summary reports will appear in scientific periodicals and will also be available in various forms as technical reports from the NISS in Research Triangle Park, North Carolina. In particular, study papers from the workshop will be prepared that will serve as indicators of further research directions, as well as current summaries of the complex issue of combining environmental data. Potential applications and improvements in associated areas of scientific/statistical research include census sampling, geostatistics, and biological effect modeling.

This workshop was an experiment in how to stimulate and foster research and collaborations across disciplinary lines. Its motivation derives, however, from ever-growing social, political, economic, and scientific needs; with such strong background, it is hoped that the workshop stimulus will be strong, compelling, and fruitful.

References

1. Krieger AM, Pfeffermann D. Maximum likelihood estimation from complex sample surveys. Surv Meth 18:225–239(1992).
2. Kish L. Statistical design for research. New York:John Wiley and Sons, 1987.
3. Overton JM, Young TC, Overton WS. Using "found" data to augment a probability sample: procedure and a case study. Environ Monit Assess 26:65–83(1993).
4. Mastala Z, Balogh KV, Salanki J. Reliability of heavy metal pollution monitoring utilizing aquatic animals versus statistical evaluation methods. Arch Environ Contam Toxicol 23:476–483(1992).
5. Vanhonacker WR, Price LJ. Using meta-analysis results in Bayesian updating: the empty-cell problem. J Bus Econ Stat 10:427–435(1992).
6. Wolf FM. Meta-analysis: quantitative methods for research synthesis. Newbury Park, CA:Sage Publications, 1986.
7. Hasselblad V, Eddy DM, Kotchmar DJ.

Synthesis of environmental evidence: nitrogen dioxide epidemiology studies. J Air Waste Manag Assoc 42:662–671(1992).
8. Greenland S, Salvan A. Bias in the one-step method for pooling study results. Stat Med 9:247–252(1990).
9. Chalmers TC. Problems induced by meta-analysis. Stat Med 10:971–980(1991).
10. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, FL:Academic Press, 1985.
11. Cox DR. Some remarks on overdispersion. Biometrika 70:269–274(1983).
12. Diggle PJ, Donnelly JB. A selected bibliography on the analysis of repeated measurements and related areas. Aust J Stat 31:183–193 (1989).
13. Carr GJ, Chi EM. Analysis of variance for repeated measures data: a generalized estimating equations approach. Stat Med 11:1033–1040(1992).
14. Austin JT, Wolfle LM. Annotated bibliography of structural equation modelling: technical work. Br J Math Stat Psychol 44:93–152 (1991).
15. Berger JO. Statistical decision theory and Bayesian analysis. New York:Springer-Verlag, 1985.
16. DuMouchel WM, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species (with discussion). J Am Stat Assoc 77:293–313(1983).
17. Carroll RJ, Ruppert D. Transformation and weighting in regression. New York:Chapman and Hall, 1988.
18. Portnoy S. Formal Bayes estimation with application to a random effects model. Ann Math Stat 42:1379–1402(1971).
19. Rubin DB, Computational aspects of analysing random effects/longitudinal models. Stat Med 11:1809–1821(1992).
20. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. Biometrics 40:961–971 (1984).
21. Reinsel GC. Mean squared error properties of empirical Bayes estimators in a multivariate random effects general linear model. J Am Stat Assoc 80:642–650(1985).
22. Schall R. Estimation in generalized linear models with random effects. Biometrika 78:719–727(1991).
23. Albert JA. Nuisance parameters and the use of exploratory graphical methods in a Bayesian analysis. Am Stat 43:191–196(1989).
24. Miller J N. Outliers in experimental data and their treatment—tutorial review. Analyst 118:455–461(1993).
25. Unwin A. How interactive graphics will revolutionize statistical practice. Statistician 41:365–369(1992).
26. Wilkinson L. Graphical displays. Stat Meth Med Res 1:3–25(1992).
27. Hosmer DW, Lemeshow S. Applied logistic regression. New York:Wiley, 1989.
28. Green RH. Sampling design and statistical methods for environmental biologists. New York:Wiley 1979.
29. Portier CP, Hoel DG. Optimal design of the chronic animal bioassay. J Toxicol Environ Health 12:1–19(1983).
30. Lewtas J, Claxton LD, Rosenkranz HS, Schuetzle D, Shelby MD, Matsushita H, Würgler FE, Zimmermann FK, Löfroth G, May WE, Krewski D, Matsushima T, Ohnishi Y, Gopalan HNG, Sarin R, Becking GC Design and implementation of a collaborative study of the mutagenicity of complex mixtures in *Salmonella typhimurium*. Mutat Res 276:3–9(1992).
31. Narula SC, Wellington JF. The minimum sum of absolute errors regression: a state of the art survey. Int Stat Rev 50:317–326 (1982).
32. Chen S, Farnsworth D. Median polish and modified procedure. Stat Prob Lett 9:51–57 (1990).
33. Cressie NAC. The origins of kriging. Math Geol 22:239–252(1990).
34. Myers DE. Interpolation and estimation with spatially located data. Chemomet Intel Lab Syst 11:209–228 (1991).